BIPARTITE GRAPH ALGORITHM WITH REFERENCE FRAME REPRESENTATION FOR PROTEIN TERTIARY STRUCTURE MATCHING

by

FAZILAH OTHMAN

Thesis submitted in fulfilment of the requirements
for the Degree of
Doctor of Philosophy

September 2010

ACKNOWLEDGEMENTS

Alhamdulillah, praise be to Allah for His wills that gave me the strength and patience to complete this thesis. I thankfully acknowledge my supervisor who is also my idol, Prof. Dr. Rosni Abdullah for her guidance, inspiration and moral support. Thank you to Assoc. Prof. Dr. Nur'Aini Abdul Rashid for your support as a teacher, colleague and friend. To Pn. Zalila Ali and Dr. Jamaludin Ali from the School of Mathematical Sciences, USM, thanks for the patience in helping me with some of the mathematical concepts. Thank you to Prof. Ulises Cortes from Universitat Politechnica de Catalunya (UPC), Spain for the opportunity to pursue four months of research attachment at UPC and the access to MareNostrum Supercomputer at Barcelona Supercomputing Center. I am grateful that I have been given the chance to utilise the computer resources, technical expertise and assistance provided by the lecturers and staff at the School of Computer Sciences, USM. Without all these supports, completing this thesis will be twice as hard. I would like to acknowledge USM and MOSTI for providing financial supports in my research works through Science Fund and Short-term grants.

My deepest gratitude to my Parlimail group, relatives, closest friends, Ibrahim, Ali Kattan, Adib, Hussein and colleagues at the PDCC Lab for their support, sharing of information and friendly care. I believe that all our hard work and efforts in gaining knowledge will benefit us in a good way. Special thanks to my beloved parents En. Othman Jaafar and Pn. Faridah Kasim and my darling siblings: Fazil, Fadli, Farhana and Firdaus for their continuous prayers, encouragement, precious love, support and tolerance, emotionally and physically during the preparation of this thesis. Thank you for always being there.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BFS	Breadth-First Search
CATH	Class, Architecture, Topology and Homologous superfamily classification database
CCSD	Complex Carbohydrate Structure Database
DFS	Depth-First Search
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FSSP	Families of Structurally Similar Proteins
GHA	Geometric Hashing Algorithm
GHARF	Geometric Hashing Algorithm with Reference Frame
MCS	Maximum Common Subgraph
NMR	Nuclear Magnetic Resonance Spectroscopy
PDB	The Brookhaven Protein Data Bank
PDB ID	Protein Data Bank Identification Number
PROTEP	PROtein Topography Exploration Programs
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristics
SCOP	Structural Classification of Proteins Database
SSAP	Structure and Sequence Alignment Program
SSE	Secondary Structure Element
TESS	TEmplate Search and Superposition program
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VAST	Vector Alignment Search Tool
WBGMRF	Weighted Bipartite Graph Matching algorithm with Reference Frame

ALGORITMA GRAF DWIBAHAGIAN DENGAN PERWAKILAN RANGKA RUJUKAN UNTUK PEMADANAN STRUKTUR TERTIER PROTEIN

ABSTRAK

Protein dengan kesamaan struktur cenderung berkongsi fungsi biologi yang sama. Ini menunjukkan kepentingan pemadanan struktur protein dalam menentukan fungsi protein. Pemadanan antara struktur baru dan struktur sasaran yang diketahui fungsinya dapat menemukan sasaran terbaik dengan persamaan tertinggi sebagai penunjuk kepada fungsi struktur baru. Penyelidikan sebelum ini menunjukkan pemadanan struktur memerlukan perkomputeran yang intensif dan memakai ruang ingatan yang besar. Masalah kekompleksan ruang diselesaikan dengan pelaksanaan dua algoritma: algoritma cincangan geometri dengan rangka rujukan (GHARF) dan algoritma graf dwibahagian berpemberat dengan rangka rujukan (WBGMRF). Dalam GHARF, perwakilan baru untuk tulang belakang protein C_α direka menggunakan rangka rujukan 3D yang berasal daripada penglihatan komputer untuk perwakilan objek. Eksperimen menunjukkan rangka rujukan sesuai untuk struktur protein. Namun, kekompleksan ruang yang tinggi kerana penggunaan jadual cincangan dalam GHARF tidak wajar untuk set data besar. Teknik ini hanya berkesan untuk protein bersaiz kecil. Dalam WBGMRF, rangka rujukan digabungkan dengan teknik pemadanan graf dwibahagian. Ujian korelasi mendapati WBGMRF mempunyai korelasi yang signifikan secara statistik dan biologi dengan program tanda aras daripada Institut Bioinformatik Eropah. Kekompleksan ruang WBGMRF menurun kepada $O(N^2+N)$ berbanding $O(N^3)$ dalam GHARF. Aplikasi rangka rujukan dalam WBGMRF dapat mengurangkan kekompleksan ruang serta memperbaiki algoritma untuk memadankan set data protein yang lebih besar dalam tempoh masa dan ruang ingatan yang munasabah.

BIPARTITE GRAPH ALGORITHM WITH REFERENCE FRAME REPRESENTATION FOR PROTEIN TERTIARY STRUCTURE MATCHING

ABSTRACT

Proteins with structural resemblances tend to share similarities in biological functions. This shows the importance of protein structure matching in function determination. Matching between a new structure and a list of target structures with known functions can discover the best target with highest similarity score to indicate the function of the new structure. Previous works have shown that structural matching is computationally intensive and consumes large amount of memory. The problem of space complexity is solved with the implementations of two algorithms: Geometric Hashing Algorithm with Reference Frame (GHARF) and Weighted Bipartite Graph Matching with Reference Frame (WBGMRF). In GHARF, new structural representation for protein C_α backbone is designed using 3D reference frame which was originally introduced in computer vision for object representation. The experiments have shown the suitability of reference frame for protein structures. Yet, high space complexity due to hash table utilisation in GHARF is unfeasible for larger datasets. Besides, this technique is only effective for matching small proteins. Hence, in WBGMRF, the reference frame is combined with bipartite graph matching technique. A correlation test shows that WBGMRF has a statistically and biologically significant correlation with the benchmark program from the European Bioinformatics Institute. The space complexity for WBGMRF is reduced to $O(N^2+N)$ compared to $O(N^3)$ in GHARF. The application of reference frames in WBGMRF reduces the space complexity and improves the algorithm for matching larger protein dataset in reasonable time and space.

CHAPTER 1

INTRODUCTION

1.1 Introduction

The advancement of high throughput machines available in DNA sequencing projects today has contributed to the massive amounts of protein sequence data. These sequences are beneficial for protein function determination which turns out to be a focal point in Structural Biology for its encouraging contributions in drug discovery and drug design (Rigden, 2009). However, amino acids sequence does not provide enough molecular information to thoroughly interpret the biological function. It is more robust to determine protein function when the protein is coiled into its 3D conformation because evolutionary origin is more apparent and preserved in protein structural properties. In view of this, there have been many research efforts on protein structure prediction to bridge the gap between protein sequence and function by offering structure prediction of available protein sequences.

Since the determinations of biological functions rely very much on the structural similarities (Krissinel, 2007), structural matching can be employed to look for resemblances in structural properties to identify the biological functions of particular proteins. Matching at structural level produces more significant structural knowledge as compared to sequence level because:

i) In molecular biology, proteins with similar structures typically have the same function (Bergeron, 2002; Westhead et al., 2002; Marti-Renom et al., 2003).

ii) During cell evolution, structure is better conserved than sequence and it is essential for function determination since macromolecules (proteins and

nucleic acids in particular) carry out most of the functions of cells (Smith, 1994).

Consequently, matching can help in protein structure prediction and furnishes information like function relationships, evolutions, and common building blocks - motifs in protein analysis. Nevertheless, protein structure matching is always associated with intensive computing and high complexity processes as its application normally involved with enormous data volume and large size proteins. This circumstance invites more research in finding a matching technique give that an accurate matching result and at the same time has an adequate space and computational complexity. The research in this thesis revolves around using a theoretical graph-based technique for protein tertiary structure matching.

1.2 Motivation

The underlying challenges in structure matching vary from the questions of data representation, matching algorithm and scoring scheme for similarity measure (May, 1999). Data representation is about how to transform all the important features of a raw data into a computable format appropriate for an algorithm or procedure. A matching algorithm will search for structural similarities that exist in the data representation which is measured using a suitable scoring scheme.

Dealing with these challenges is important in protein matching as it involves complex structures that carry along geometrical properties such as atomic coordinates, dihedral angle and fold characteristics. Thus, having suitable structural representations that allow reliable distinction between different objects in the database with the ability to resolve firmly on the matching result and overall computation time is a crucial task.

Structure matching can be conducted in different ways at different levels and constraints. There are several methods to perform structure matching and one of the prominent ways is using alignment. Nevertheless, alignment between the query and target structure in 3D space requires a calculation of rotation and translation position of the structure (Nussinov & Wolfson, 1991; Wang et al., 2004) which is known to be time consuming and computationally intensive (Taylor & Orengo, 1989; Zuker & Somorjai, 1989; Aung et al., 2003).

Since data representation and matching algorithm are two interrelated elements in matching, it is motivating to design a representation that is invariant to rotation and translation and to apply it on matching algorithm that is unconstrained by rotation and translation steps, but it still able to produce a good matching result. By leaving out the rotation and translation steps, it is anticipated that the space complexity can be reduced and the computational complexity of the matching algorithm can be improved. Indirectly, this characteristic prepares the matching algorithm for dataset with massive amount of structural data because protein in nature is large and normally available in a high volume database.

Protein structure matching is a computationally intensive, time-consuming and memory space-consuming process. It is important for a matching technique to have a significant integration between efficient structural data representation and matching algorithm to trim down the space and time consumption in the procedure. These observable facts motivated this research towards finding a good data representation for protein tertiary structure and matching algorithm that can produce good matching result and to reduce the space and computational complexities.

1.3 Research Questions and Research Objectives

The approach to solve this problem is by looking into the similar problems in other fields. Image processing and object recognition in computer vision are the closest studies that can relate to computational biology, particularly with the significant roles of graphs in computer vision (Shokoufandeh & Dickinson, 2002). The images and protein structures share a clear characteristic in term of the data structure where the image points and protein atoms can be defined using the Cartesian coordinates x, y and z in 3D space (Fischer et al., 1994). Since there have been extensive research in computer vision, rather than in biology, it is intriguing to explore more on graph for this area (Zaslavskiy, 2010).

The objectives of this work are identified by answering these research questions:

i) What is a reliable representation to represent geometrical properties of protein tertiary structure?

ii) Is the representation suitable to keep the structural information for matching?

iii) Can the representation improve the matching accuracy?

iv) Can the combination of different graph representation and matching algorithm improve the accuracy and reduce the space complexity?

v) Can the combined algorithm provide an optimum matching result?

The research objectives are:

i) To study and to propose a graph-based representation that is suitable for protein backbone features,

ii) To propose an efficient graph-based matching algorithm for protein structure and apply graph-based data representation identified in (i) to the algorithm,

iii) To design a workflow and to develop a framework for protein tertiary structure matching, and

iv) To evaluate the efficiency of the algorithm and the accuracy of the matching results.

1.4 Approach and Research Methodology

In order to achieve the research objectives stated in Section 1.3, graph theoretic approach is proposed for the data representation and algorithm for matching towards solving structural matching problem. The decision is made based on these justifications:

i) Graph matching technique can find a maximum matching that is in line with the aim to find matched structures with higher similarity.

ii) Combination of graph–based matching technique with a graph-based reference frame representation will be a good combination to solve matching problem.

iii) Graph-matching technique is independent to rotation, translation or alignment operations. Escaping this operation is expected to reduce the computational intensity of the algorithm.

The outline of the main framework for this research is illustrated in Figure 1.1. Protein structural data in PDB file format are retrieved from Protein Data Bank (PDB) (Berman et al., 2003) by referring to the PDB identification number (PDB ID) of each structure. In pre-processing phase, these files, or which considered as raw protein data will go through a cleaning and extraction processes before it is stored in a local database. Prior to matching, the pre-processed data is transformed into a suitable data representation in reference frame construction phase. In structure

matching and evaluation, matching algorithm will compare protein structures represented in reference frames. The reported results will be evaluated based on the performance metrics and comparisons with a benchmark program.

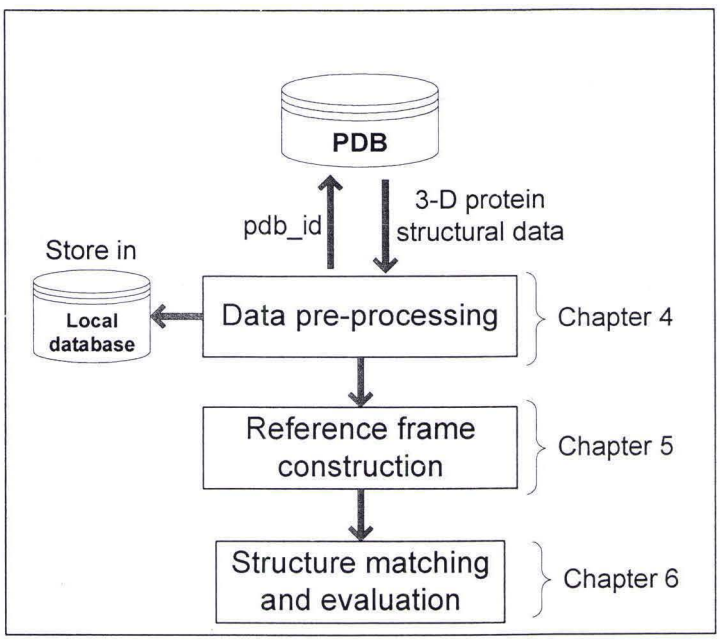

Figure 1.1: The outline of the main framework.

1.5 Scope and Limitation

This research focuses on protein tertiary structure matching. Although the underlying challenges are data representation, matching algorithm and similarity measure, the anticipated contributions only touch on data representation and algorithm for matching. For similarity function to measure the correspondences in matching, an existing function is deployed from one of the resources in related work. Creating another similarity function is another branch in research. In terms of data representation, a full profile of protein tertiary structure is considered, where only C_α atoms are extracted from protein backbone. As for matching algorithm, the experiments to test the matching algorithm are carried out on a pre-processed dataset

containing structures of the same size so as to avoid gap handling which is required in handling dataset with unequal structure size.

1.6 Contributions

Protein structure matching enables discovery of structural properties embedded in protein that holds a strong evolutionary trait to preserve its biological function. This thesis presents a graph-based matching using reference frame as data representation to represent protein tertiary structure and weighted bipartite graph matching algorithm to find structural similarity.

1.6.1 Reference Frame as Data Representation

First approach focuses on one of the main concerns of any application in computational biology that is data representation. 2D reference frame has been originally introduced in computer vision to represent 2D objects. In this work, the 2D reference frame is modified to suit the protein tertiary data. Instead of two points, now three points (or three protein atoms) are needed to calculate the three orthogonal vectors for the 3D reference frame. The adaptation of 2D reference frame into a 3D reference frame to suit the proposed protein tertiary structural data is as follows:

A reference frame is described as a platform or plane that works as a reference to define a coordinate system (Eidhammer et al., 2004). On an object with 2D image points (x-y coordinate), a reference frame contains two orthogonal vectors calculated from a pair of points. A unit vector from a point at coordinates (0, 0) to point (1, 0), or also known as basis pair is defined as x–axis. The orthogonal vector perpendicular to a basis pair is set as the y–axis as illustrated in Figure 1.2a. Once a reference frame has been created, the remaining points in the object will be transformed based on this

reference frame (Eidhammer et al., 2000). A set of newly transformed points operate as one of the object instances described in the particular coordinate system.

As a modification to problems involving 3D points (x-y-z coordinate), three points are needed to create a 3D reference frame as illustrated in Figure 1.2b. Assume that three imaginary points plotted at coordinates (0, 0, 0), (0, 1, 0) and (1, 0, 0) will create a plane. The orthogonal vector from point (0, 0, 0) to point (0, 1, 0) is declared as x-axis. Second vector from point (0, 0, 0) to point (1, 0, 0) is needed to calculate an orthogonal vector pointing at 90° from the plane as the y-axis. Next, the vectors that work as x-axis and y-axis are used to calculate the next orthogonal vector to be the z-axis. In total, three orthogonal vectors are calculated from a triplet of points to construct the 3D reference frame.

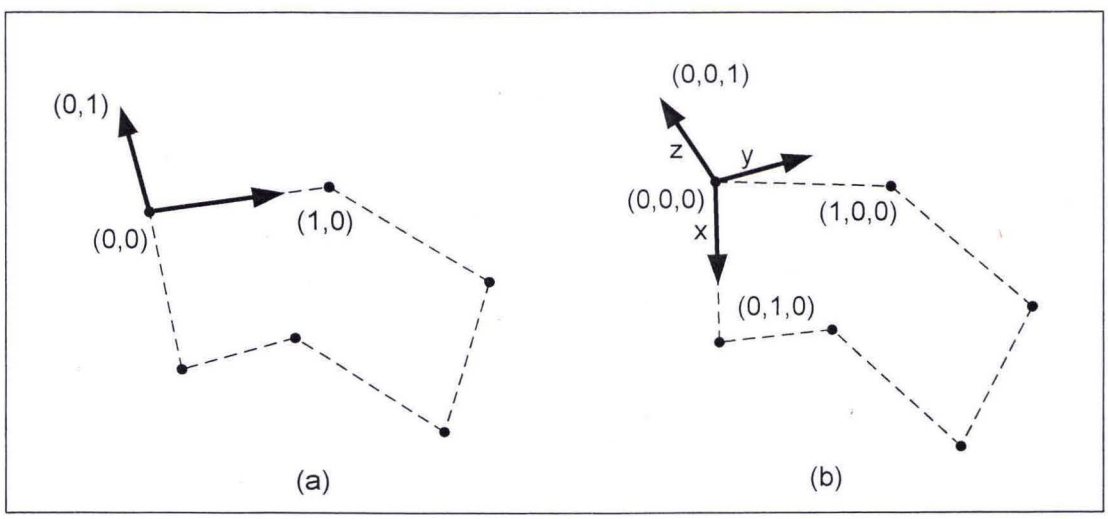

Figure 1.2: (a) 2D reference frame with two orthogonal vectors. (b) 3D reference frame with three orthogonal vectors constructed from triplet of points.

To test the workability for this newly designed reference frame on tertiary structure, it is experimented with geometric hashing algorithm (GHA) which is known for its primary use with reference frame in computer vision. The experimental results show that the proposed reference frame can represent and conserve the

geometrical features in protein, yet the limitations are caused by its fundamental application in GHA that has high space complexity for the hash table use, and it only performs best on dataset of small proteins. Combination of reference frame with other matching algorithm may open a good prospect of utilising the reference frame to its maximum capability.

1.6.2 Bipartite Graph Matching Algorithm with Reference Frame Representation

In this contribution, the proposed reference frame is retained and a new integration with bipartite graph based matching technique is suggested. To do this, the bipartite graph matching is designed to be adapted to the reference frame. Two partitions on bipartite graph represent two separate structures. The vertices of each partition denote a reference frame calculated from triplet of atoms, thus the amount of vertices in each partition depends on how many reference frames have been created from each structure. Another feature to distinguish the proposed bipartite graph approach with others is that, each vertex will contain matching vectors of the respective reference frame. An edge is drawn between two vertices to denote a matching between vertices of separate partition. The edge is weighted with the similarity score calculated between two matching vectors using an exponential similarity measure. After the graph construction, the matching is computed using weighted bipartite matching algorithm. The algorithm starts by looking for initial matching, and then Ford-Fulkerson algorithm with breadth first search (BFS) are used to find augmenting path in the graph until it reaches maximum matching.

The common way of matching is usually using basic distance measure between single matching vectors from two structures. Here, a new way of matching with

regards to the integration of reference frame with bipartite graph is suggested. In matching, all matching vectors in query structure and all matching vectors in target structure (the two partitions) are taken into account. Then, the correspondences with maximum weight that can be fitted in every match are investigated. Matching with maximum weight is reported as the matching result.

The quantitative performance comparison with benchmark program (the Secondary Structure Matching program (SSM)) shows that the proposed method is statistically significant and outperforms the application of reference frame with GHA. Furthermore, the utilisation of reference frame has been successfully maximised. In its former application with GHA, reference frame is best applied on dataset of small proteins. But, the new integration with bipartite graph, reference frame becomes useful for matching on dataset with large proteins.

1.7 Thesis Organisation

The organisation of this thesis is guided by Figure 1.1.

Chapter 2 presents preliminary knowledge for the research. It gives an introduction to the proteins building block starting from DNA as the ground unit to protein in its primary sequence, secondary structure, tertiary structure and quaternary structure. The terminologies and concepts in protein structure matching are briefly explained to give preliminary understanding about the application of matching in the area of computational biology. The definition and components of bipartite graph is also presented in this chapter.

Chapter 3 reviews on structure matching technique and discusses the underlying issues in matching. A broad literature review about matching technique and data

representation will be presented before concentrating into the existing graph-based matching techniques and structural representations that have been proposed not only in the area of structural biology but also in other fields.

Chapter 4 covers the research methodology. The description covers the details of pre-processing phase and the evaluation of the experimental matching results. The pre-processing phase is about the preparation of the experimental data. It starts with selection and retrieval of protein data from PDB, how to clean the raw data from erroneous records and the extraction of atomic records seize for matching. The matching results will be evaluated using statistical methods: correlation coefficient test and performance metrics such as sensitivity, specificity and accuracy. The explanation embraces the results' benchmark with SSM and the utilisation of SCOP classification as a standard to determine the correct matching.

Chapter 5 dedicates to the first contribution in this thesis that is the implementation of geometric hashing algorithm with reference frame (GHARF). It describes the construction of reference frame from backbone atoms C_α as data representation for protein tertiary structure. The design is accomplished with its application with GHA. The experimental results on two protein datasets are presented and discussed at the end of the chapter.

Chapter 6 presents the second contribution: the implementation of weighted bipartite graph matching with reference frame (WBGMRF). The main workflow is given and explained and discussed throughout the chapter. It covers the construction of the bipartite graph and the combination of reference frame with bipartite graph matching technique. Although there is no new similarity measure, but rather to

deploy an existing similarity function, the adaptation to the matching problem is still be mentioned in this chapter.

Chapter 7 comprises the conclusion and suggestions of possible improvements for future work. The findings and achievements in accordance to the research objectives are concluded here.

CHAPTER 2

BACKGROUND

2.1 Introduction

Bioinformatics is a new scientific discipline that unites specialisation from the fields of biology, computer sciences and mathematical sciences in order to understand the living system. Figure 2.1 shows a fusion of different fields towards bioinformatics. Generally, bioinformatics problems have the tendency of having large data volumes, intrinsic complexities and require additional methods for dealing with error prone data. Due to these characteristics, solutions to these problems demand large memory spaces and intensive computing.

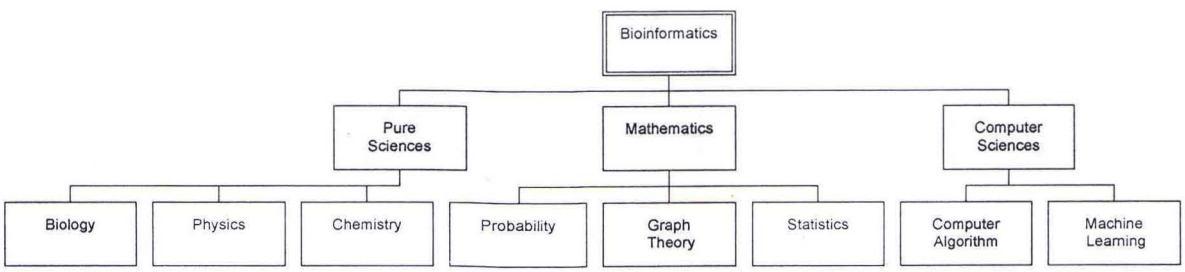

Figure 2.1: Combination of different disciplines such as pure sciences, mathematics and computer sciences in bioinformatics.

For example, swift acquisition of biological data stimulates the needs for huge data storage and high performance computing for data processing. Data processing, most of the time, requires the same group of instructions to be repeated trillions times to accomplish the given task. If these tasks are going to be executed in a conventional manner, it will certainly delay the processing time or it is almost impossible to be executed. Therefore, to fulfil the demands mentioned above, computer sciences come into the picture for its capacity to provide computing power

and large memory space. As a result, the whole complete process can be performed in faster processing time and extend the capability of the algorithm to handle larger datasets. These situations motivate computer scientists to explore the research in bioinformatics.

There are quite a number of bioinformatics applications that have been actively explored in computer sciences such as structure prediction, image processing, sequence analysis, data filtering, data mining, pattern recognition and structure matching. Among all these applications, this research concentrates on protein structure prediction and structure matching.

Section 2.2 and Section 2.3 start with the building blocks of biological data, starting from the simplest DNA unit, until protein tertiary structure. After the familiarity of the data, Section 2.4 covers the experimental and computational methods for protein structure prediction and then highlights on the role of matching operations in the computational methods. Section 2.5 defines structure matching in the context of protein structure followed by Section 2.6 that gives the basic concepts of graph including the definition and matching components of a specific graph type called bipartite graph. The theoretical concept of graph-based matching algorithm is also covered in this section.

2.2 DNA as the Smallest Unit

Deoxyribonucleic acid (DNA) is a seed of biological data which is acquired from DNA genome sequencing projects in sequence form. DNA has four types of chemical bases which are adenine (A), guanine (G), cytosine (C), and thymine (T), where structurally, A is paired with T, and G is paired with C (Figure 2.2). These pairing units of A-T and C-G are known as base pairs.

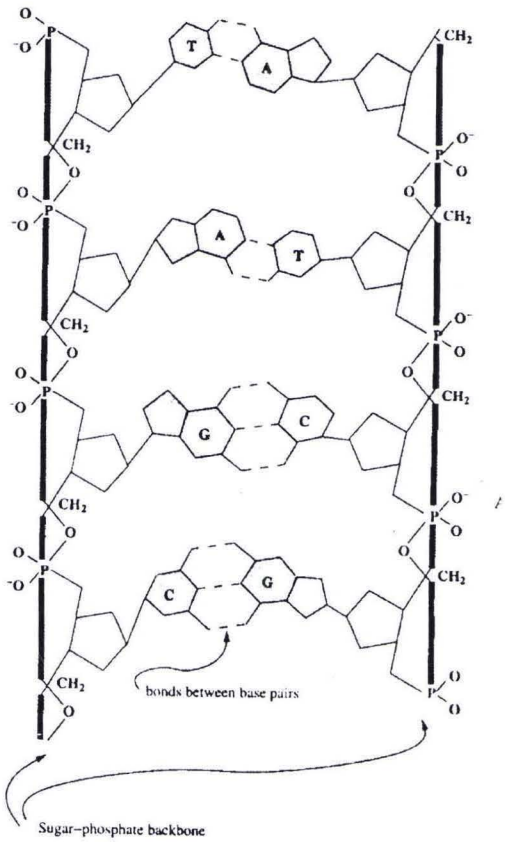

Figure 2.2: Illustration of pairing units in DNA (Eidhammer et al., 2004).

A DNA sequence can be translated into a protein sequence by transforming the arrangements of base triplets into one amino acid or protein residue. These arrangements of base triplets are known as codon (Eidhammer et al., 2004). The detailed arrangements of codons and protein residues can be referred to the universal genetic code in Table 2.1. Protein sequences become meaningful in function determination when they are transformed into two dimensional (2D) and three dimensional (3D) structures (Chien-Cheng et al., 2004). Function determination which is the main goal in structural biology, is nonetheless an important research to deliver structural knowledge for rational drug design, protein re-engineering and protein bio-molecule interactions (Smith, 1994).

Table 2.1: The universal genetic code table (Eidhammer et al., 2004)

First position	Second position				Third position
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

2.3 Protein Structure

Protein is one of the biological macromolecules that are vital to support cell activities in living organism. Each protein has its own biological function for examples Keratin for the growth of hair, nails and skin, and Actin and Myosin for the development of muscle tissues in the body. Protein structure consists of amino acids that are connected to each other by one or more polypeptides chains. There are 20 different amino acids that take different combinations and different lengths to form a protein. These 20 amino acids which can be denoted using single alphabet or three alphabets are shown in Table 2.2. Amino acid denoted by single alphabet is known as a residue.

Table 2.2: Names of amino acids and its respective one-letter code and three-letter code (Eidhammer et al., 2004)

Amino acid	One-letter code (residue)	Three-letter code
Alanine	A	Ala
Cysteine	C	Cys
Aspartic acid	D	Asp
Glutamic acid	E	Glu
Phenylalanine	F	Phe
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Lysine	K	Lys
Luecine	L	Leu
Methionine	M	Met
Asparagine	N	Asn
Proline	P	Pro
Glutamine	Q	Gln
Arginine	R	Arg
Serine	S	Ser
Threonine	T	Thr
Valine	V	Val
Tryptophan	W	Trp
Tyrosine	Y	Tyr

The chemical structures of amino acids contain the exact components of amino group ($-NH_2$) and carboxyl group (-COOH) but differs in terms of the side-chain substances (*–R*) as illustrated in Figure 2.3. The side-chains are unique between different amino acids. For examples in Figure 2.4, Glycine has the simplest side-chain with only a single hydrogen atom (H), Serine has an alcohol (-OH) side-chain and Alanine has hydrocarbon side-chain (containing only hydrogen and carbon) (Eidhammer et al., 2004).

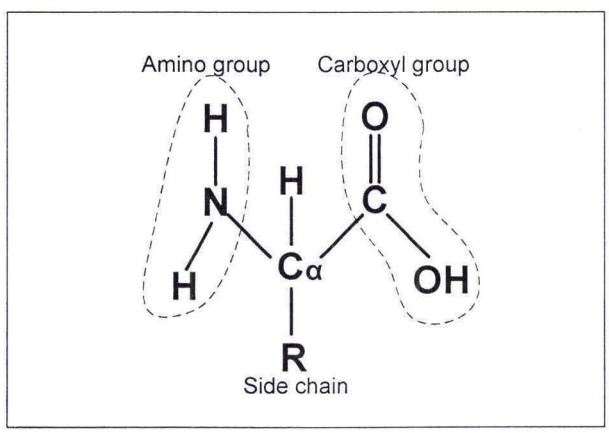

Figure 2.3: Chemical structure of a single amino acid.

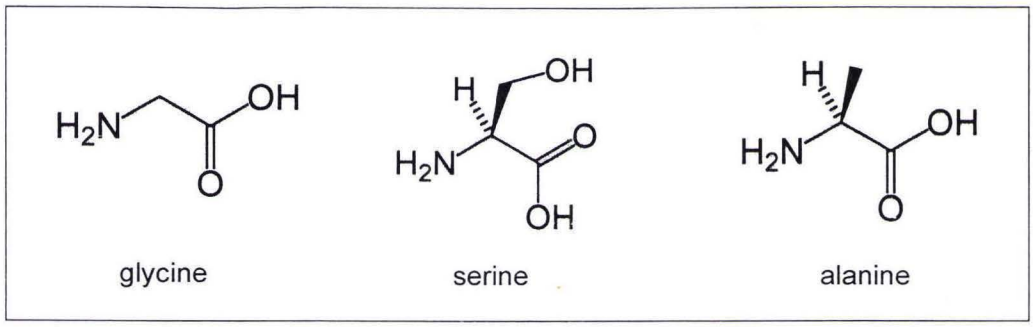

Figure 2.4: Chemical components for Glycine, Serine and Alanine.

A protein molecule is formed from a chain of amino acids sequence that folds into complex 3D structures such as primary, secondary, tertiary and quaternary structural levels (Bergeron, 2002). Each structural level has its own approach to describe the qualities hidden in a protein. The diagram for each structural level is shown in Figure 2.5. The primary structural level refers to the order of amino acids sequence along a polypeptide chain. From the primary structures, the formation of secondary structure elements (SSEs) such as beta-sheets (or β-sheets or β-strands) and alpha-helices (α-helices) can occur by hydrogen bonds between carboxyl and amino group. The arrangements of these SSEs describe the secondary structural level for the protein. When a single unit of SSE is attracted to other SSEs, they are connected by regions called loops, turns or coils to construct a tertiary structure. A protein tertiary structure describes a shape of a protein that folds due to the attractions between SSEs in a single polypeptide chain. A quaternary structure is formed when several proteins with separate polypeptide chains are packed together. As an example, the functional hemoglobin consists of two alpha-hemoglobin proteins and two beta-hemoglobin proteins.

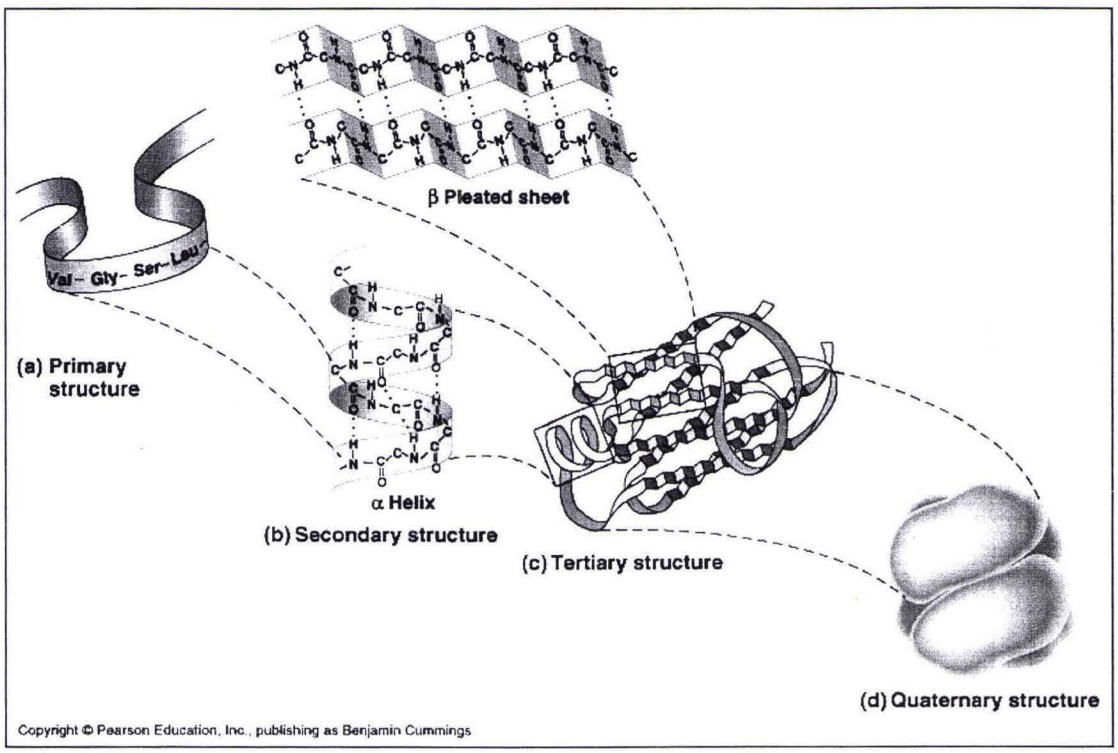

Figure 2.5: The arrangements of protein in its sequence, secondary structure, tertiary structure and quaternary structure (Campbell & Reece, 2002).

In the context of protein function determination, 3D structures hold more functional qualities compared to the information in amino acid sequence (Krissinel & Henrick, 2004b). For instance, the earliest work in structure alignment by Perutz et al. (1960) has discovered that myoglobin and hemoglobin that store and transport oxygen in blood are similar in structures and functions, but different in their amino acid sequences. Hence, the properties that allow us to determine protein function are more preserved when the protein is coiled into its 3D conformational structure.

Since this thesis focuses on structure matching at tertiary level, the discussion will focus more on the tertiary structure. Every amino acid or residue in proteins contains atoms as the smallest unit which is accompanied with 3D atomic Cartesian coordinates. These atomic coordinates are identified experimentally and can be retrieved from PDB, an archive of experimentally determined 3D biological

macromolecules. Protein backbone fragment can be constructed in two ways. First is from chain of backbone atoms alpha-Carbon (C_α) and second from chain of Nitrogen-alpha Carbon-Carbon (N-C_α-C) atoms from linked amino acids as illustrated in Figure 2.6. Although these two options are different, implementing either of them provides enough information to obtain protein backbone (Bartoli et al., 2008). Nevertheless, generating protein backbone from C_α atoms has been frequently applied in protein structure prediction (Havel, 1998). Therefore, protein structure matching based on backbone fragment is acceptable as the backbone construction is not changeable or moveable (Bergeron, 2002; Westhead et al., 2002).

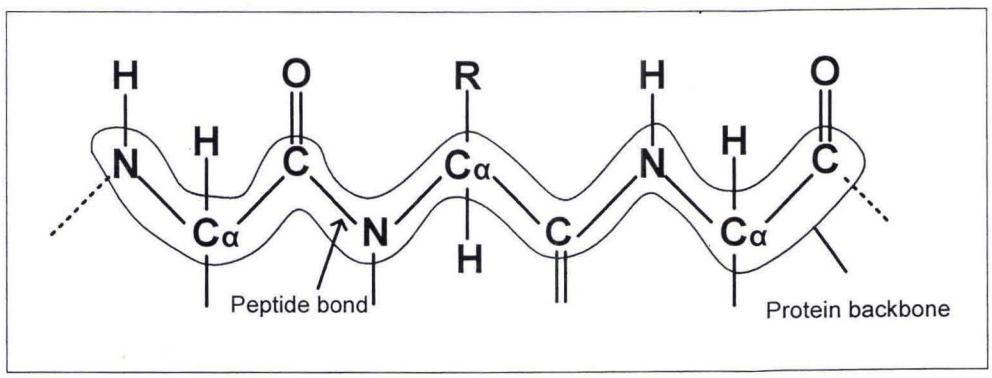

Figure 2.6: Protein backbone fragment from a chain of amino acids.

2.4 Protein Structure Prediction Methods

In molecular biology, discoveries in protein functions become the central focus in research as they provide answers to protein re-engineering and drug discovery. Although protein sequences are important in many studies, 3D structures hold strong traits to protein function since it can only be performed when the macromolecules are coiled into its specific 3D shapes (Smith, 1994). Furthermore, the bonds between proteins structure and protein functions are supported by these two facts:

Fact 1: Proteins exhibiting related functions are likely to share some similarities in their 3D structures (Chien-Cheng et al., 2004; Marti-Renom et al., 2003; Westhead et al., 2002; Bergeron, 2002; Huang et al., 2005; Wu et al., 2007; Krissinel, 2007; Chen & Chen, 2003).

Fact 2: Proteins with similar 3D structures often have related functions even if their 1D amino acid sequences are not alike. 3D structure conserved more functional properties than sequence (Chien-Cheng et al., 2004; Marti-Renom et al., 2003).

Before any protein function determination is carried out, the structure has to be predicted first. There are two types of protein structure prediction methods:

i) Experimental prediction method.

ii) Computational prediction method.

2.4.1 Experimental Prediction Methods

X-ray crystallography and multidimensional nuclear magnetic resonance (NMR) spectroscopy are examples of experimental methods for predicting protein tertiary structure. Table 2.3 describes the pros and cons of these experimental prediction methods. Although these methods are more accurate compared to computational methods, unfortunately, they are considered as difficult, time consuming and expensive. These experimental limitations have prevented these methods from keeping pace with the increasing number of protein sequences (Orengo et al., 2003). X-ray crystallography method for instance is limited by the difficulties in obtaining diffraction-quality crystal forms and isomorphic heavy atom derivatives, whereas NMR spectroscopy requires the proteins to be dissolved in a very concentrated solution and has limitations on the size of proteins that are

amendable to study. These limitations make it unlikely for the rate of experimental structure determination to accelerate dramatically in the near future.

Table 2.3: Description of experimental protein structure prediction methods

Methods	Description
X-ray crystallography / X-ray diffraction	▪ Accurate, but time consuming and expensive. ▪ Must be able to crystallize protein. ▪ Require large amount of material (~20mg).
Multidimensional NMR Spectroscopy	▪ Accurate, but time consuming and expensive. ▪ Limitation on size of protein (only 120 residues). ▪ Protein must be soluble ~ 30mg /ml. ▪ Able to locate flexible and rigid regions.

PDB archive assembles protein tertiary structures predicted using experimental method such as X-ray crystallography and multidimensional NMR from the community worldwide. Referring to RCSB PDB Annual Report in July 2008 (Research Collaboratory for Structural Bioinformatics, July 2008), there was a rapid growth in PDB with 20 new structures received daily, and accumulated 50,000 structures were added to the archive in April 2008. With incoming structures actively identified using high throughput machines, an efficient technique is needed to classify structural properties between new and available protein structures. This is where the computational method becomes more practical.

2.4.2 Computational Prediction Methods

Computational methods can be divided into three distinctive techniques:

i) Comparative modelling (homology modelling),

ii) Fold recognition, and

iii) *ab initio* (*de novo*).

2.4.2(a) Comparative Modelling

Comparative modelling looks into homologous sequences which can be retrieved from available protein databases and it uses existing information such as family classification, protein structure details, sequence and dihedral angles values for protein structural prediction (Dayalan et al., 2004; Marti-Renom et al., 2000, 2002). In this technique, homologous sequences are substantial to show that the sequences have agreed to a certain level of similarity and share a common biological evolution. For example, similar sequences with more than 20% of identity are likely to have a similar structure which later may impose similar structure function. MODBASE (Pieper et al., 2006) and Modeller (Eswar et al., 2007) are examples of the automated systems that employ this technique.

Figure 2.7 shows a standard framework of comparative modelling as highlighted in Marti-Renom et al. (2003). Steps in comparative modelling can be divided into four main processes: fold identification and template selection, target-template alignment, model building and assessment of models.

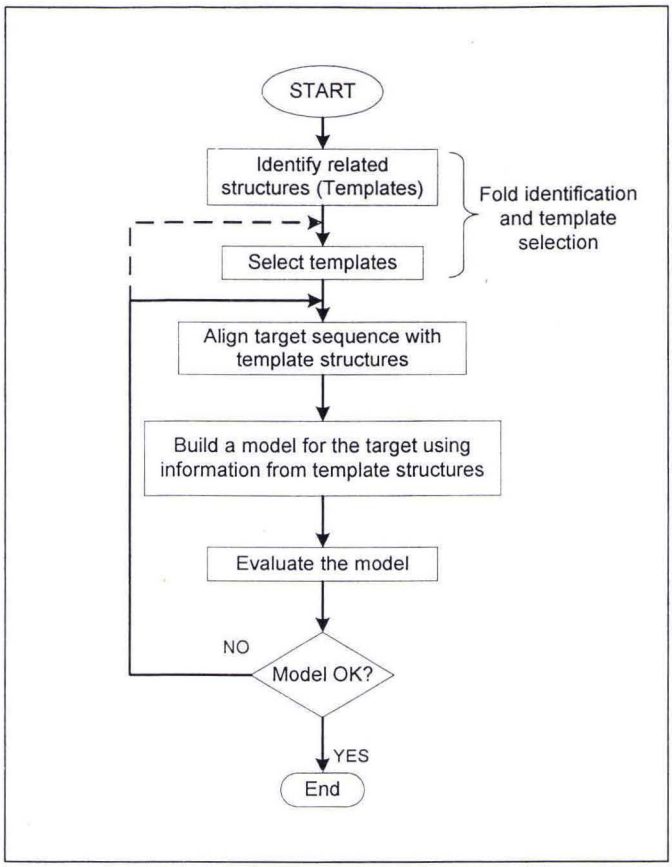

Figure 2.7: Standard framework for comparative modelling (Marti-Renom et al., 2003).

i) Fold Identification and Template Selection

Given a target sequence with unknown structure, the template structures are selected by searching for sequences of known structure from existing databases such as PDB, SCOP (Murzin et al., 1995) or CATH (Orengo et al., 2002), that is homologous to the target sequence. After that, fold identification method is used to identify similar features such as fold comprising between target sequence and templates. These features will be the foundation to form protein model for target sequence. Fold identification methods can be divided into three categories:

- Pairwise sequence-sequence comparison such as BLAST,
- Multiple sequence comparisons such as PSI-BLAST, and
- Threading or 3D template matching for example 3D-PSSM program.